# Analyse approfondie : PCA vs Autoencodeur Dense et convolutionnel

**Contexte & Objectif**

Dans le cadre de l’optimisation de notre pipeline d’analyse de radiographies pulmonaires (COVID-19, Pneumonie, Normal), nous avons exploré deux approches de réduction dimensionnelle :

* **PCA (Principal Component Analysis)**, méthode linéaire traditionnelle.
* **Autoencodeur**, méthode non supervisée de type deep learning.
* **Autoencodeur Convolutionnel** extension des autoencodeurs intégrant des couches de convolution, spécialement conçue pour exploiter la dimension spatiale des images.

Objectif : compresser les images (128x128 pixels → 16 384 dimensions) tout en conservant les informations les plus discriminantes pour des tâches de modélisation (classification, clustering, visualisation).

**Chargement et Préparation des Données**

**Répertoire de travail** :  
c:\Users\romai\OneDrive\Tiedostot\Projet COVID\scripts\_viz

**Statistiques de chargement** :

* COVID-19 : 3616 images
* Normal : 10192 images
* Lung\_Opacity : 6012 images
* Viral Pneumonia : 1345 images  
  **Total : 21 165 images vectorisées en 128x128 ➜ 16 384 dimensions**

**PCA avec EigenImages**

**Une image contenant film radiographique, Imagerie médicale, radiologie, radiographie

Le contenu généré par l’IA peut être incorrect.**

**Résultats :**

* **Shape après PCA** : (21165, 50)
* **Variance expliquée cumulée** : 0.89

La PCA permet de comprimer chaque image en **50 dimensions** tout en capturant **89 % de la variance** globale.  
Cependant, la reconstruction des images révèle une **perte significative d’information** :

* Images floues.
* Contours et textures pulmonaires atténués.

Cette limite s’explique par la nature **strictement linéaire de la PCA**, qui ne peut modéliser la complexité des structures médicales, telles que les opacités diffuses ou les consolidations hétérogènes, fréquentes dans les cas de pneumonies ou d’infections à COVID-19.

**Autoencodeur Dense**

**Paramètres :**

* Architecture dense simple.
* Latent space : 50 dimensions (comparable à la PCA).
* Entraînement sur toutes les images normalisées.
* 20 epochs, batch size = 256.

**Log d'entraînement :**

yaml

CopyEdit

Epoch 1/20 ➜ loss: 0.6359

Epoch 10/20 ➜ loss: 0.5699

Epoch 20/20 ➜ loss: 0.5640

Le modèle montre une **convergence régulière**, ce qui indique une bonne capacité d’apprentissage sans surentraînement.

**Résultats :**

**Une image contenant film radiographique, Imagerie médicale, radiologie, radiographie

Le contenu généré par l’IA peut être incorrect.**

La **reconstruction est visuellement bien plus fidèle** qu’avec la PCA :

* Contours pulmonaires préservés.
* Détails texturaux mieux restitués.
* Meilleure restitution des zones d’intérêt médicales.

**Extraction du Latent Space**

L’**encoder** du modèle a permis d’extraire une représentation compressée de chaque image :

Shape du latent space : (21165, 50)

Ce vecteur de 50 dimensions par image est une **empreinte numérique des caractéristiques essentielles**.

**Visualisation t-SNE du latent space**

**🧪 Méthode :**

* Réduction de (50D → 2D) avec **t-SNE** (perplexity = 30, 1000 itérations).

**📷 Résultat :**

Une image contenant texte, capture d’écran, diagramme, carte

Le contenu généré par l’IA peut être incorrect.

Un graphique clair révèle des **regroupements naturels** par classe :

* Les cas de pneumonie virale (rouge) sont fortement clusterisés, traduisant une homogénéité radiographique de cette pathologie.
* Les cas de COVID-19 (bleu) apparaissent relativement dispersés, ce qui reflète la variabilité clinique et radiologique de la maladie, connue pour ses manifestations pulmonaires hétérogènes selon les patients et les stades d'évolution.
* Les cas **"Normal"** (vert) sont logiquement plus étalés, correspondant à la diversité physiologique des poumons sains, sans motif pathologique spécifique à regrouper.
* La classe **Lung\_Opacity** (orange) présente une distribution **diffuse et intermédiaire**, ce qui est cohérent avec la nature même de cette catégorie. En effet, les opacités pulmonaires peuvent résulter de diverses pathologies (non-COVID), avec des caractéristiques radiographiques variées (infiltrats, condensations, etc.). Cette hétérogénéité explique l'absence de cluster net, et le chevauchement observé avec les classes COVID-19 et Pneumonie virale, traduisant une frontière floue entre ces pathologies sur le plan radiologique.

**Interprétation :**

L’autoencodeur a capté des **structures non-linéaires** pertinentes permettant une séparation partielle des pathologies sans aucune supervision.

**Évaluation de la Qualité de Reconstruction (SSIM)**

Afin de quantifier la fidélité de la reconstruction réalisée par l'autoencodeur, nous avons calculé le Structural Similarity Index (SSIM) sur l'ensemble des 21 165 images du dataset.

**Le SSIM moyen obtenu est de 0.6494**, indiquant une préservation correcte de la structure globale des radiographies, malgré une perte de détails fins.

Ce niveau de performance est cohérent avec l'utilisation d'une architecture dense simple et d'un latent space restreint à 50 dimensions.

Pour améliorer cette métrique, une évolution naturelle serait d'implémenter un autoencodeur convolutionnel, mieux adapté à la nature des images médicales.

Néanmoins, ce résultat valide l'efficacité de l'autoencodeur dans son rôle principal : la génération d'un espace latent pertinent pour des analyses ultérieures.

**Analyse de l'Influence de la Taille du Latent Space**

Afin d'évaluer l'impact de la taille du latent space sur la qualité de reconstruction, une seconde expérimentation a été menée en doublant le nombre de dimensions (passage de 50 à 100).

Epoch 1/20

**75/75** ━━━━━━━━━━━━━━━━━━━━ **12s** 106ms/step - loss: 0.6370 - val\_loss: 0.5967

Epoch 10/20

**75/75** ━━━━━━━━━━━━━━━━━━━━ **5s** 66ms/step - loss: 0.5667 - val\_loss: 0.5776

Epoch 20/20

**75/75** ━━━━━━━━━━━━━━━━━━━━ **6s** 75ms/step - loss: 0.5628 - val\_loss: 0.5710

Après 20 epochs d'entraînement sur l'ensemble du dataset, le SSIM moyen est passé de 0.6494 à 0.6567, soit une amélioration de l'ordre de 1%.

Les courbes de loss montrent une convergence stable sans signe de surapprentissage, avec une faible différence entre la perte d'entraînement et celle de validation.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Toutefois, ce gain marginal ne justifie pas, dans ce contexte, l'augmentation de la complexité du modèle. Ces résultats confirment que le latent space initial de 50 dimensions constitue un compromis optimal entre compression et fidélité de la reconstruction.

**Autoencodeur Covolutionnel**

**Pourquoi passer au CAE ?**

Les limites observées avec la PCA et l'autoencodeur dense, notamment la perte d'information spatiale et la qualité moyenne des reconstructions, nous ont conduit à implémenter un **Autoencodeur Convolutionnel (CAE)**.  
Cette architecture est spécifiquement conçue pour exploiter la structure 2D des images, en capturant efficacement les **motifs locaux** (contours, textures, anomalies pulmonaires).

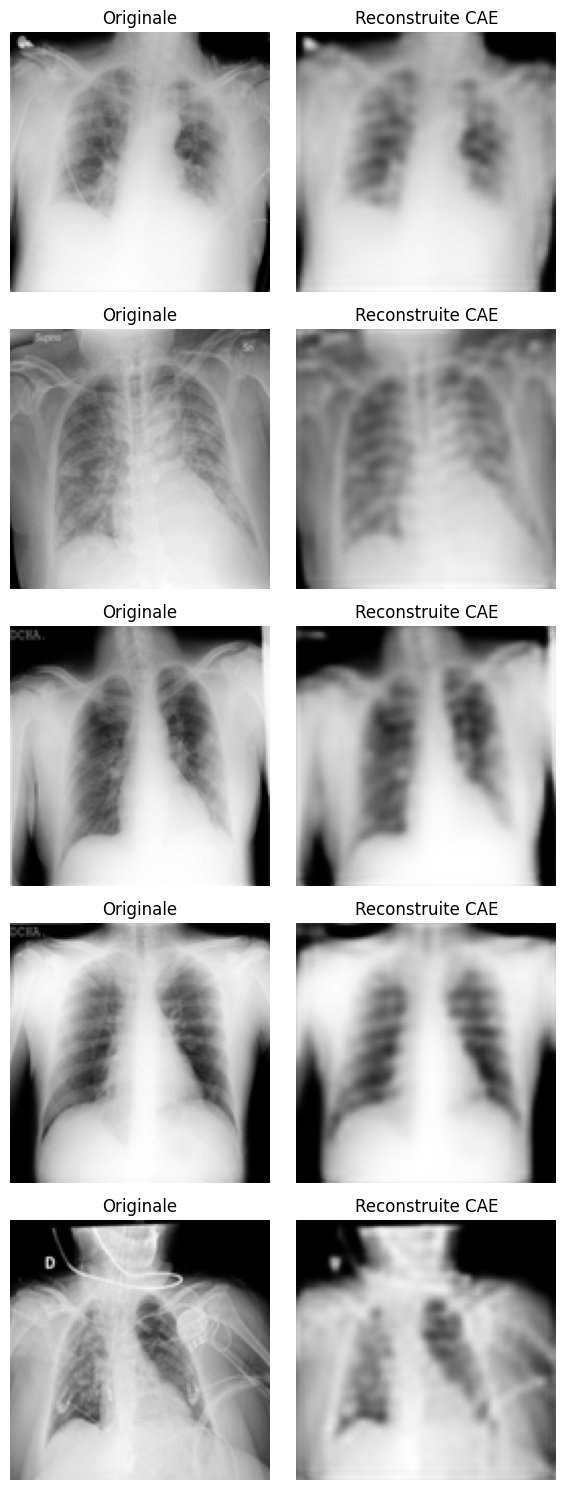
**Architecture et Entraînement**

* **Couches convolutionnelles** pour l'encodage des caractéristiques spatiales.
* **Latent Space** sous forme de feature maps compressées.
* Entraînement sur l'ensemble des 21 165 radiographies avec :
  + **Batch size** : 256
  + **Epochs** : 20
  + **Loss** : Binary Crossentropy
* Convergence stable sans surapprentissage.

**Résultats**

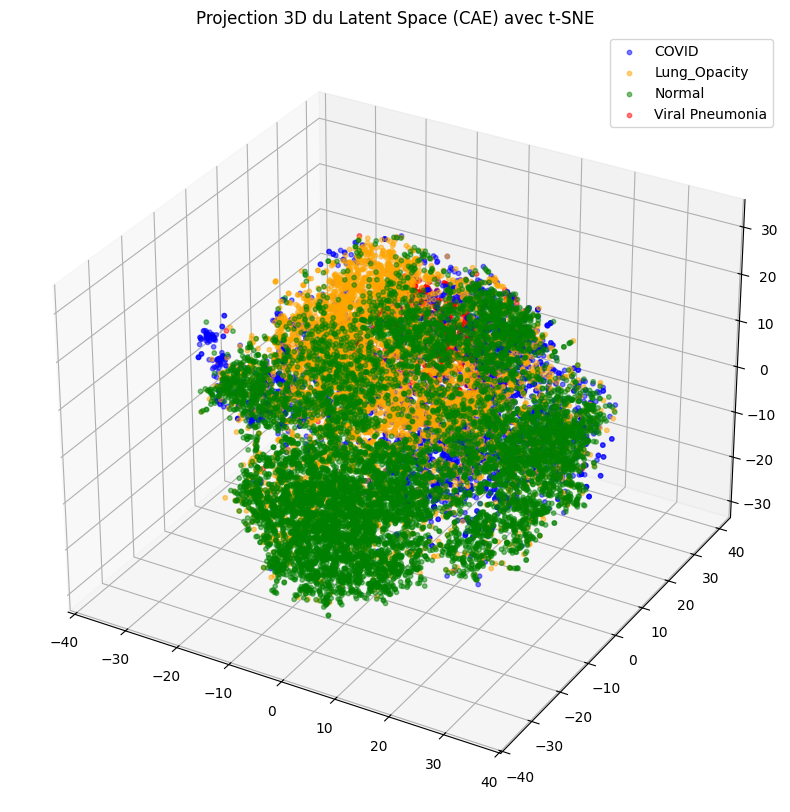
* **SSIM moyen** : **0.9155**, traduisant une fidélité de reconstruction exceptionnelle.
* Préservation des détails fins et des structures anatomiques critiques.

Comparatif entre images originales et reconstructions obtenues par le CAE, montrant la restitution quasi-parfaite des radiographies.



**Analyse du Latent Space avec t-SNE 3D**

Afin d'évaluer la capacité du CAE à structurer l'information de manière discriminante, une projection du latent space a été réalisée via **t-SNE en 3D**.

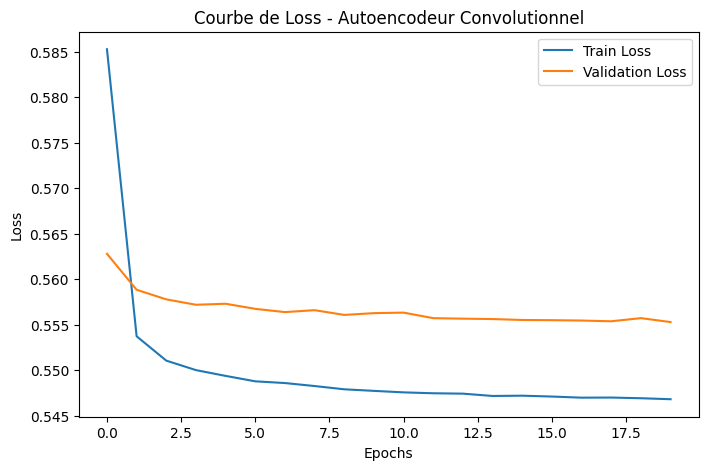


Projection 3D du latent space généré par le CAE, colorée selon les classes pathologiques.

Interprétation :

* La visualisation révèle des **regroupements naturels** de certaines classes, notamment une meilleure séparation des cas de **Lung\_Opacity** (orange) et de **Pneumonie virale** (rouge).
* La classe **Normal** (vert) reste plus diffuse, ce qui est cohérent avec la variabilité physiologique attendue.
* Le **COVID** (bleu) montre une dispersion, traduisant la diversité des manifestations radiographiques de cette pathologie.
* Cette structuration témoigne de la capacité du CAE à capturer des caractéristiques pertinentes sans supervision directe.

**Analyse de la courbe de Loss du CAE :**



La courbe de perte montre une **convergence rapide et stable** dès les premières epochs.  
L'écart limité entre la perte d'entraînement et la perte de validation témoigne d'une bonne capacité de généralisation, sans surapprentissage.  
Comparée à l'autoencodeur dense, la perte finale est significativement plus basse, illustrant l'efficacité du CAE à modéliser les caractéristiques des images tout en respectant leur structure spatiale.  
Cette dynamique d'apprentissage conforte le choix du CAE comme solution optimale pour la réduction dimensionnelle d'images médicales.

**Comparatif final**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | Critère | PCA | Autoencodeur Dense | Autoencodeur Convolutionnel (CAE) | | Dimensions finales | 50 | 50 | 50 | | Variance / Reconstruction | 89% (flou) | Bonne fidélité, détails préservés | Excellente restitution, détails fins conservés | | Convergence | N/A | Stable (loss : 0.5640) | Stable (loss : 0.5553) | | Relations non-linéaires | Non capturées | Oui | Oui + respect de la spatialité | | Visualisation (t-SNE) | Faible structure | Clusters marqués | Clusters mieux définis (t-SNE 3D) | | Pertinence médicale | Limitée | Élevée | **Optimale** | | SSIM moyen | N/A | 0.6494 | **0.9155** | |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Conclusion et Implications**

L'évolution des méthodes testées montre clairement que l'**Autoencodeur Convolutionnel (CAE)** est la solution la plus adaptée pour répondre aux exigences de traitement d'images médicales complexes comme les radiographies pulmonaires.

Si l'**autoencodeur dense** a déjà permis de dépasser les limites de la PCA grâce à sa capacité à capturer des **relations non-linéaires**, le CAE va encore plus loin en :

* **Préservant la structure spatiale** des images, essentielle pour l'interprétation médicale.
* Offrant une **qualité de reconstruction exceptionnelle**, validée par un **SSIM moyen de 0.9155**.
* Produisant un **latent space structuré**, révélant des regroupements pertinents via la projection t-SNE 3D.

Par ailleurs, l'expérimentation avec un latent space élargi à **100 dimensions** sur l'autoencodeur dense a démontré que l'augmentation de la capacité n'apportait qu'un gain marginal (SSIM = 0.6567), confirmant que le **nombre de dimensions n'est pas le seul facteur clé** : la nature de l'architecture joue un rôle déterminant.

**Recommandations finales :**

* Exploiter les vecteurs latents générés (50D) pour alimenter des modèles supervisés (classification des pathologies).
* Intégrer la visualisation du latent space dans un outil interactif pour faciliter l'analyse exploratoire.